CALL FOR PROPOSALS

EXCHANGE SCHEMES

OC3-2025-TES-01

ENFIELD: EUROPEAN LIGHTHOUSE TO MANIFEST TRUSTWORTHY AND GREEN AI

Co-funded by the European Union

TABLE OF CONTENTS

Introduction

This is the catalogue of challenges available to the second out of four open calls for individual researchers exchange under the ENFIELD (European Lighthouse to Manifest Trustworthy and Green AI) project, co-funded by the European Union. Through the ENFIELD Exchange Scheme open calls and the Financial Support to Third Parties (FSTP) mechanism, the project aims to attract the top-level researchers to conduct foundational research activities related to specific scientific/technological challenges in artificial intelligence, contributing to ENFIELD network creation and expansion to European AI labs.

Pillar Green-AI

G-AI.1 Green AI Metrics

Keywords: Green AI Metrics; Computational Cost; Energy Usage Monitoring; Energy Efficiency; Sustainability.

STATE-OF-THE-ART

State of the art in monitoring Green AI metrics is focused on developing standardized, accurate metrics to measure the environmental impact of AI throughout its lifecycle. These metrics aim to evaluate AI architectures not only for performance accuracy but also for energy efficiency and reduced carbon emissions, including hardware manufacturing impacts. Significant work is directed at estimating the computational efficiency, like floating-point operations (FLOPs), for various AI models, facilitating comparisons under fixed computational budgets, crucial for SMEs with limited resources. In industry, there's a movement towards integrating Green AI principles into system development to promote efficiency and robustness without relying solely on the latest hardware advances. The widespread adoption of generative AI, including Large Language Models (LLMs), has further underscored the urgency of addressing energy consumption due to their massive computational requirements. This trend highlights the critical importance of developing and adopting Green AI metrics to monitor, report, and optimize energy use, while also enabling resource-efficient solutions in low-power environments such as IoT devices and remote deployments.

SCIENTIFIC CHALLENGES

One of the primary scientific challenges in monitoring Green AI metrics lies in the establishment and standardization of these metrics across varied AI system architectures. This includes not only the computation of the efficiency and accuracy of algorithms but also accounting for the environmental impact throughout the AI lifecycle, from hardware production to operational deployment.

Reliability of these metrics is another key challenge ensuring that measurements are consistent, reproducible, and representative of real-world usage. Transparency in how these metrics are calculated is crucial to prevent misinterpretation or misuse. Researchers must grapple with the dearth of universally accepted metrics and the limited availability of comprehensive data needed to assess the full environmental footprint of AI technologies. This challenge is compounded by the need for interdisciplinary collaboration to ensure that any developed metrics are both technically sound and environmentally meaningful.

One of the goals of this challenge is to create a suite of standardized Green AI metrics that balances performance with energy efficiency, guiding the design of AI systems that are both robust and sustainable. For a comprehensive proposal, researchers would require access to current AI models, energy consumption data, and cross-sectoral environmental impact assessments, alongside the tools for the simulation and evaluation of AI architectures under these new metrics.

RESEARCH ACTIVITIES

Multiple tools exist today and can be used to estimate in real-time the carbon emissions of training and using AI models. However, these tools are often limited to specific contexts, leaving significant gaps in coverage and adaptability. The research activities here should focus on filling the gaps by providing new tools in contexts that were not considered before. At the code level, existing tools are generally Python libraries that can be used when developing AI-based algorithms. There is a need for new real-time tools, programmed in different and more low-level languages such as C/C++. Such tools could enable broader compatibility and performance in resource-constrained environments. There is also a need for tools that can be deployed on a wider range of platforms, such as edge devices and non-standard computing systems. The research activities may include the development of C/C++ tools/libraries dedicated to providing different metrics related to CO2 emissions and electricity usage, when training and using AI models. Focus is suggested to be on deep learning algorithms and industrial scenarios, involving robotics and machine vision, such as waste sorting applications or disassembly operations. Additionally, the research must address the accuracy, scalability, and interpretability of these tools, ensuring they are not only technically robust but also accessible for practical deployment. Contrary to LLMs and models applied to cross-section data, there is a gap in measuring the emissions of time series algorithms. Particularly, most of the competitions that assess the performance of these algorithms (e.g., Makridakis competitions) just focus on prediction capabilities (sometimes in both, point forecasts and prediction intervals). Therefore, there is a need to include computational efficiency metrics in these competitions. We consider that metrics mixing forecasting power and efficiency (low computational cost - emissions) should be developed and tested, so that time series models could be compared and ranked in these two dimensions.

EXPECTED RESULTS

The research results will lead to the development of new tools and new benchmarks, that could be beneficial for anyone developing AI-based algorithms. The ENFIELD project aims to provide to the scientific community novel tools, metrics, and guidance to develop greener algorithms in various use cases. These results are expected to be disseminated through published papers, open-source code, and targeted dissemination activities, ensuring widespread accessibility and impact.

This challenge is not just about reducing computational costs, it's about reshaping AI's role in addressing global sustainability challenges. By developing robust, standardized Green AI metrics, we can guide the next generation of AI technologies towards more environmentally responsible innovation.

POSSIBLE HOST ORGANISATIONS

- UCM (Alfredo Garcia-Hiernaux)
- DTI (Francois Picard)

G-AI.2 Physics-Informed Machine Learning

Keywords: Physics-Informed Machine Learning; Bayesian Modeling; Causal Models; Physics-Informed Neural Networks.

STATE-OF-THE-ART

Physics-informed machine learning utilizes insights derived from (physical) theories during the development and training of machine learning models. Its potential to reduce the required amount of training data/duration and model complexity has a direct influence on the resulting model's environmental footprint. Physics-informed machine learning (such as physics-informed neural networks, physics-guided neural networks, but also feature engineering and the inclusion of architecture constraints) thus not only improves generalization and extrapolation properties but can also contribute to making AI greener.

SCIENTIFIC CHALLENGES

(Physical) Theories representing prior knowledge must be formalized in a way that can be exploited during model development and/or training.
The ideal approaches to incorporate knowledge into AI systems (such as feature engineering, regularization, and architecture selection) must be adapted to the considered application.
The trade-offs between energy consumption, model complexity, and accuracy achieved with physics-informed ML are not always obvious and rarely quantified.

RESEARCH ACTIVITIES

Physics-Informed Machine Learning (PIML) approaches show significant potential for enhancing the performance of learning models, particularly when the application can be partially or fully represented by a physical model. To test and validate this potential, PREDICT proposes to provide (at a minimum) a dataset derived from its activities in the energy and/or manufacturing sectors, as well as the physical model associated (partial or complete) to the considered system, for the application of physics-informed ML to solve a concrete task (e.g., estimating the power of an electrical network, predicting the quality of a manufactured part, etc.). This requires identifying relevant physical theories, incorporating them during model development/training, and evaluating model performance. The resulting model shall further be compared to purely data-driven and purely mechanistic models in terms of runtime, energy consumption, and accuracy.

EXPECTED RESULTS

By collaborating with researchers from KNOW and PREDICT, the project seeks to develop ML models that minimize environmental impact. Expected outcomes include a description of the methodology adopted for the integration of the physical model in the learning stage, the overall framework describing the utilisation of the PINN/PIML, as well as the models for selected tasks in relevant application domains that are characterized by small energy consumption and runtime, and that require small amounts of training data. A further expected result is an improved understanding of the positive effects of physics-informed ML. Results of the project shall be published at an international academic venue.

POSSIBLE HOST ORGANISATIONS

Know Center Research GmbH (Franz Rohrhofer, Bernhard Geiger)
Note: this challenge is co-hosted by PREDICT

Pillar Green-AI

G-AI.3 The Policy Landscape for Green AI

Keywords: Green AI; Sustainable AI; AI Act; Electronics Supply Chains; Critical Materials Act; AI Policy; AI Regulation.

STATE-OF-THE-ART

Investment in and use of generative AI has grown significantly in recent years and is only poised to grow more. Evidence shows that this sector is already creating vast amounts of e-waste, contributing the global growth of e-waste production at a rate that is five times faster than the rate of growth for recycling programs. Additionally, AI hardware is produced as part of global electronics supply chains, which are known to cause extensive environmental damage, with downstream effects on human health and safety. Existing regulations for AI, namely the EU's "AI Act", make reference to respecting societal and environmental well-being and accountability, but the act focuses on regulating the design and use of AI models in terms of their direct applications. In terms of regulating the environmental impact of AI, including the computational energy cost, physical waste streams, and AI hardware supply chains, it appears that the AI Act may have a rather large blind spot. This raises the question of whether other EU legislation is applicable to regulating AI, like the EU Supply Chain Act or the Critical Raw Materials Acts. At present, there appears to be no overlap nor interaction between these various acts and questions remain as to how and to what extent these various acts may effectively regulate AI hardware production, operations, and waste streams. What further complicates the regulatory landscape is the disconnect between the consumption of AI-based services by European citizens and the geographic location of service providers (e.g., ChatGPT), and the global nature of hardware supply chains.

SCIENTIFIC CHALLENGES

- Identify the European policy and regulatory interactions and overlaps that might apply to a robust understanding of Green AI, and the policy and regulatory blind spots and potential loopholes.
- Identify key stakeholders shaping the European policy and regulatory landscape for Green AI.
- Identify the key challenges for the creation of a robust European environmental policy and regulation for AI systems.

RESEARCH ACTIVITIES

- Development of a comprehensive European policy and regulatory overview and analysis for Green AI that includes AI operations, deployment, hardware production, and end-of-life.
- Provide a socio-political analysis and stakeholder mapping of key European policy, regulatory, industry, environmental and civil society stakeholder relationships, power dynamics, and leverage points.
- Develop an environmental policy/regulatory recommendation for AI.

EXPECTED RESULTS

- Innovative and state-of-the-art policy and regulatory overview.
- A preprint and subsequent journal publication that reports the research undertaken and the findings.
- A policy brief that can be disseminated on behalf of ENFIELD.

POSSIBLE HOST ORGANISATIONS

- Know Center Research GmbH (Nicki Lisa Cole)
- Telenor Research & Innovation (Jeriek Van den Abeele)

Pillar Green-AI

G-AI.4 Green Generative Language Models

Keywords: Generative AI; Large Language Models; Energy Consumption; Environmental Impact; Predictive Process Monitoring.

STATE-OF-THE-ART

Advancements in large language models (LLMs) have revolutionized natural language understanding and generation, but their substantial computational demands raise environmental concerns. Current research focuses on enhancing LLM energy efficiency through techniques like model pruning, quantization, and distillation, alongside optimizing training processes with efficient hardware and renewable energy, as well as improving fine-tuning and prompting techniques, e.g., leveraging in-context learning (ICL). Edge computing is also being explored to distribute processing loads, reducing reliance on energy-intensive data centres. These efforts aim to create powerful yet environmentally sustainable LLMs. On the application side, using pre-trained LLMs with ICL could also reduce energy consumption by adapting to diverse tasks without costly retraining.

SCIENTIFIC CHALLENGES

The challenge focuses on making generative language models and their deployment greener by optimizing with regards to both their energy consumption and performance. Researchers will analyse the energy consumption of various LLMs to identify patterns and inefficiencies, and investigate mitigation methods. Key scientific challenges include developing criteria for selecting energy-efficient LLMs for specific tasks without compromising performance, managing temporal credit assignment for delayed feedback, and aligning AI objectives with user goals to avoid suboptimal outcomes. The research may also explore techniques like model pruning, quantization, and distillation to reduce computational load. Barriers include the complexity of accurately measuring energy use in diverse environments, the scarcity of high-quality data, and ethical concerns like privacy and bias. Researchers will access existing LLMs, energy-efficient computing resources, and datasets to evaluate energy consumption and performance, enabling the development of innovative, sustainable methodologies for deploying LLMs for various tasks. One application of interest is predictive process monitoring (PPM), where LLMs can analyse patterns in event log data from business processes like order tracking.

RESEARCH ACTIVITIES

- **Energy consumption analysis:** Conduct comprehensive analyses of energy consumption patterns across various LLMs to identify inefficiencies and areas for improvement.
- **Energy mitigation strategies**: Explore and implement techniques such as model pruning, quantization, and distillation to reduce energy consumption while preserving performance standards.
- **Optimization of LLM selection criteria**: Develop and refine criteria for selecting the most energy-efficient LLMs tailored to specific tasks, ensuring optimal performance without excessive energy use.
- **Analysis of trade-offs with fine-tuning/ICL**: Studying trade-offs between LLM performance and energy consumption in different fine-tuning and ICL setups for predictive process monitoring.
- **Temporal credit assignment management**: Investigate and address challenges related to temporal credit assignment for delayed feedback, enhancing the sustainability of LLM learning processes.
- **AI objective alignment**: Develop methodologies to align AI objectives with user goals, ensuring efficient and effective interactions between AI systems and human users, avoiding suboptimal outcomes.
- **Ethical considerations**: Seamlessly incorporate ethical considerations, such as privacy and bias, into the development and deployment of sustainable LLM methodologies to ensure responsible AI use.

EXPECTED RESULTS

The ENFIELD project aims to achieve significant scientific progress in the development of greener generative language models and understanding energy consumption and performance trade-offs in their applications. Expected outcomes include the establishment of energy-efficient methodologies for analysing and optimizing LLMs, novel techniques for mitigating energy consumption in LLMs, criteria for selecting energy-efficient models, insights on energy-efficient fine-tuning/ICL setups, and strategies for aligning AI objectives with user goals. The expected results from the exchange include publications in scientific conferences or journals. Ultimately, the project seeks to advance the state of the art in sustainable AI by providing actionable insights and methodologies for deploying greener generative language models in real-world applications.

POSSIBLE HOST ORGANISATIONS

- SINTEF (Erik Johannes Husom and Sagar Sen)
- Telenor Research & Innovation (Mike Riess and Jeriek Van den Abeele)

G-AI.5 Energy-Efficient Large Language Models for Sustainable Software Engineering

Keywords: Energy Efficiency; Large Language Models (LLMs); Sustainable AI; Inference Optimization; Software Engineering; Real-Time Energy Monitoring; Carbon Footprint; Predictive Energy Models; Prompting Strategies.

STATE-OF-THE-ART

Energy Efficiency, Large Language Models (LLMs), Sustainable AI, Inference Optimization, Software Engineering, Real-Time Energy Monitoring, Carbon Footprint, Predictive Energy Models, Prompting Strategies

Large Language Models (LLMs) have revolutionized software engineering by automating tasks like code synthesis, bug detection, and vulnerability analysis. However, their significant energy consumption during inference has raised concerns about cost and environmental impact. Existing efforts, such as quantization and pruning techniques, aim to improve efficiency but often overlook real-time adaptability and comprehensive monitoring. Tools like MELODI provide detailed energy metrics for inference. Despite these advancements, a holistic, multi-model framework for real-time energy monitoring and optimization remains lacking.

SCIENTIFIC CHALLENGES

- **Energy Consumption Transparency**: Accurate, real-time tracking of both CPU and GPU energy use across diverse hardware platforms and LLM architectures.
- **Scalability:** Ensuring energy-efficient operation of LLMs across varying task complexities, model sizes, and dataset characteristics.
- **Predictive Modeling**: Developing models to dynamically predict energy usage based on task-specific parameters and optimize in real-time.
- **Sustainability Metrics**: Incorporating carbon footprint and cost alongside energy efficiency to provide a holistic view.

RESEARCH ACTIVITIES

- **Energy Monitoring Framework:** Investigate tools to track energy use for both CPU and GPU in real time across multiple LLMs.
- **Optimization Strategies:** Investigate prompting techniques, token limitations, and architectural choices to improve energy efficiency.
- **Predictive Energy Models:** Train lightweight ML models to predict energy consumption based on input features, enabling dynamic optimizations.
- **Evaluation Benchmarks:** Create standardized benchmarks to compare energy efficiency across models, tasks, and hardware platforms.

EXPECTED RESULTS

- A comprehensive, adaptable framework for real-time energy monitoring of LLM inference across diverse tasks and hardware.Identification of best practices for energy-efficient prompting, token management, and model deployment strategies.
- Predictive models capable of dynamically optimizing LLM operations for energy efficiency.
- Benchmarks and tools for evaluating energy consumption and sustainability of LLMs in software engineering.
- Insights to guide future research on energy-efficient AI for software engineering applications.

POSSIBLE HOST ORGANISATIONS

- SINTEF (Arda Goknil)

G-AI.6 Green World Models

Keywords: World Model Building; Foundation Models; Energy Efficient World Model Building.

STATE-OF-THE-ART

World Models based on foundation models represent the forefront of artificial intelligence, leveraging large-scale pre-trained architectures like transformers to simulate, predict, and interact with complex environments. These models integrate diverse modalities—language, vision, and action—enabling unified representations of the world. Advances include neural-symbolic reasoning, cross-modal learning, and reinforcement learning with foundation models, yielding breakthroughs in autonomous agents, robotics, and digital twins. Key innovations focus on efficient fine-tuning, memory augmentation, and real-time adaptation. World Models excel in scalability, generalization, and zero-shot learning, providing robust frameworks for decision-making and long-horizon planning, driving progress in AI-driven simulations, virtual environments, and real-world applications.

SCIENTIFIC CHALLENGES

Building world models for digital twins using foundation models and digital twin formalisms (e.g. DEVS) faces several challenges. First, aligning data-driven foundation models with structured, rule-based formalisms is complex due to differing semantics. Temporal and causal reasoning, critical for digital twins, is challenging for foundation models, which struggle with explicit dynamics. Ensuring validation, explainability, and integration of probabilistic outputs with deterministic constructs adds complexity. Scalability is limited by computational demands, while handling uncertainty and domain-specific constraints requires hybrid solutions. Furthermore, a lack of standardized methodologies for interoperability between unstructured foundation model outputs and formal models hampers seamless integration.

RESEARCH ACTIVITIES

The project will focus on developing energy-efficient techniques for fine-tuning foundation models, designing hybrid frameworks that integrate neural and symbolic reasoning, and improving temporal and causal modelling for digital twins. Efforts will include establishing scalable architectures for handling large-scale data inputs, ensuring interoperability between models and domain-specific applications, and creating standardized methodologies for validation and explainability. By addressing these challenges, the research aims to bridge the gap between data-driven AI and structured formalisms, enabling robust, sustainable world modelling for real-world applications.

EXPECTED RESULTS

The project aims to achieve several key outcomes:

- **Efficient World Model Architectures**: Development of energy-efficient methodologies for fine-tuning and deploying foundation models for world modelling, significantly reducing computational and energy costs while maintaining accuracy.
- **Hybrid Frameworks for Digital Twins**: Creation of robust hybrid frameworks that integrate neural foundation models with structured, rule-based formalisms, enabling accurate and interpretable digital twins for complex systems.
- **Enhanced Reasoning Capabilities**: Development of advanced techniques for temporal and causal reasoning, improving the ability of world models to predict, simulate, and adapt to dynamic environments effectively.
- **Validation and Standardization**: Establishment of standardized methodologies and benchmarks for evaluating, validating, and explaining world models, ensuring trustworthiness and transparency in real-world applications.
- **Domain-Specific Applications**: Demonstration of scalable, energy-efficient world models applied to key domains such as energy management, climate impact simulation, and industrial digital twins, showcasing their potential for sustainable and impactful applications.
- **Academic Contributions**: Publication of findings in leading AI and systems engineering venues, contributing to the foundational understanding and practical advancements in world model development.

POSSIBLE HOST ORGANISATIONS

- SINTEF (Sagar Sen)

G-AI.7 Cooperative Multi-agent Green AI

Keywords: Multiagent AI; Foundation Models; Cooperation; Deep Multiagent Reinforcement Learning.

STATE-OF-THE-ART

Cooperative multiagent AI augmented by foundation models focuses on helping multiple AI systems work together effectively by using tools like debate, self-reflection, and teamwork training. In **multiagent debate**, AI systems share and challenge ideas to improve solutions. **Self-reflection** allows each system to double-check its decisions to stay on track with shared goals. Using **deep reinforcement learning**, AI learns how to cooperate in complex situations by sharing rewards and communicating better. These approaches are now being combined to tackle real-world problems like dividing resources fairly, resolving disputes, and addressing environmental challenges, making AI collaboration smarter and more practical for everyday use.

SCIENTIFIC CHALLENGES

Key challenges include optimizing energy efficiency during deployment of multiagent AI augmented by foundation models to reduce carbon footprints, ensuring sustainable coordination by minimizing communication overhead and redundant computations, and developing scalable, eco-friendly systems that align with sustainability goals. Balancing computational demands with environmental considerations remains a critical focus for greener AI development.

RESEARCH ACTIVITIES

Key efforts include creating benchmark datasets for evaluating multiagent systems, developing algorithms for efficient communication among agents, and applying these technologies to domains like manufacturing, space, health, and energy to address real-world challenges while promoting environmental sustainability and optimizing resource use.

EXPECTED RESULTS

The project aims to achieve several key outcomes:

- **Methodology and Frameworks:** Development of robust methodologies for integrating foundational models into cooperative multiagent systems, enabling improved debate, self-reflection, and teamwork. This includes a comprehensive framework describing how foundational models enhance decision-making, consensus-building, and cooperation.
- **Energy-Efficient Communication:** Development of multiagent AI models optimized for minimal energy consumption during communication. These models focus on reducing redundant exchanges, streamlining information sharing, and enhancing message efficiency, ensuring effective collaboration in real-world domains such as resource management, conflict resolution, and environmental sustainability while minimizing environmental impact.
- **Domain-Specific Solutions:** Implementation of case-specific AI solutions, showcasing how cooperative systems can address complex challenges like equitable resource allocation or climate impact mitigation effectively and sustainably.
- **Theoretical Insights:** Enhanced understanding of how foundational models and multiagent techniques, such as debate and self-reflection, contribute to better collaboration and environmental efficiency in AI systems.
- **Academic Contributions:** Publication of findings, methodologies, and case studies at international academic venues, fostering further research in cooperative multiagent AI and its applications.

POSSIBLE HOST ORGANISATIONS

- SINTEF (Sagar Sen)

Pillar Adaptive-AI

A-AI.1 Adaptive AI for Environmental Monitoring: Multimodal Data Fusion for Context Aware Deployment

Keywords: Multimodal Data Fusion; Environmental Monitoring; Data Integration; AI-based Adaptation; Temporal Modelling.

STATE-OF-THE-ART

Recent advances in adaptive AI have improved the processing of complex, multimodal data for environmental monitoring. Techniques like Vision Transformers (ViTs) and CLIP integrate satellite, radar, and geospatial data but face challenges in handling heterogeneous resolutions, frequencies, and noise. Self-supervised methods, including adaptive denoising and contrastive learning, help mitigate sensor noise and data gaps, though further refinement is needed. Temporal models, such as Transformers with domain adaptation, address seasonal and regional variability but struggle with generalization. Interpretability remains a priority, with feature attribution methods like Integrated Gradients and SHAP, albeit at a computational cost. Efficient deployment in resource-constrained settings relies on pruning and quantization. Despite progress, key challenges persist in data fusion, noise reduction, adaptability, interpretability, and efficiency.

SCIENTIFIC CHALLENGES

Recent breakthroughs in adaptive AI have highlighted its potential to address the following key challenges in Environmental Monitoring and Data Fusion: **1) Heterogeneous and multimodal data fusion**: combining data from diverse sources (e.g., sentinel-1 radar, DEM, hydrological data) requires models to learn shared embeddings while preserving modality-specific details, which is computationally complex and prone to inconsistencies. **2) Noisy and sparse datasets**: addressing speckle noise, atmospheric interference, and gaps in radar data, along with a scarcity of labelled flood data, poses challenges for training robust models with minimal supervision. **3) Temporal and spatial variability**: adapting to dynamic environmental characteristics across regions and seasons (e.g., monsoons, snowmelt, or flash floods) requires models capable of handling large-scale variability in patterns. **4) Semantic disentanglement and interpretability**: ensuring models can explain how features like elevation, precipitation, and radar reflectance contribute to flood susceptibility predictions is critical for stakeholder trust but remains difficult in complex, high-dimensional models. **5) Resource-Constrained Deployment**: Environmental Monitoring requires access to high-performance computing which flood-prone area, for example, often lack, necessitating lightweight models that balance accuracy, scalability, and efficiency.

RESEARCH ACTIVITIES

The research activities to be carried out in this project are organized as follows: **1) Data Preparation & Scoping:** Collect and pre-process multimodal environmental datasets (e.g., satellite imagery, radar, geospatial data and data from Sentinel-1, DEM, Sentinel-2, hydrology); Analyse data noise, sparsity, and variability across different modalities; Define evaluation metrics for model performance (e.g., accuracy, robustness, interpretability); **2) Baseline Model Development:** Develop and evaluate architectures for multimodal data fusion; Integrate self-supervised and unsupervised methods for noise and data gap mitigation; Incorporate lightweight optimizations like pruning and quantization to enhance efficiency; **3) Heterogeneity & Noise Mitigation:** Develop context-aware attention mechanisms for handling noisy, multimodal data; Experiment with adaptive denoising and unsupervised learning techniques to improve data quality and model robustness; **4) Temporal & Spatial Adaptation:** Incorporate temporal modelling strategies to address seasonal and regional variability in environmental data; Apply domain adaptation methods to enhance model generalization across different environmental conditions; **5) Interpretability & Efficiency:** Implement feature attribution and explainability techniques for transparent decision-making; Develop lightweight and resource-efficient models through pruning, quantization, and other optimization methods; **6) Validation & Deployment:** Conduct end-to-end testing and evaluation on real-world environmental scenarios; Refine models based on testing feedback, focusing on robustness, scalability, and user stakeholder input for deployment.

EXPECTED RESULTS

By the end of the project, a fully functional and validated adaptive multimodal model for environmental monitoring will be developed, flood monitoring and prediction will be the main focus. Key outcomes will include a cleaned, multimodal dataset and well-defined evaluation metrics, laying a solid foundation for model development. The final model will showcase effective integration of diverse data sources and data fusion, leveraging adaptive AI techniques to produce robust, noise-resilient models capable of handling sparse and noisy datasets. Through domain adaptation, the model will be applicable to a variety of environmental scenarios, with an emphasis on interpretability and efficiency to ensure a lightweight, deployable system in resource-constrained environments.

The model's performance will undergo comprehensive end-to-end testing, with the final product being a deployment-ready system. These results will be disseminated through two key scientific outputs: a conference paper focusing on data fusion strategies and noise mitigation, and a detailed journal article outlining the entire model development process, including interpretability, generalization, and deployment strategies.

A validated framework for adaptive AI in environmental monitoring, emphasizing:

- Effective integration and processing of multimodal data.
- Robust and generalizable performance under noisy and dynamic conditions.
- Lightweight, deployable solutions tailored for resource-constrained environments.

Dissemination of results will include research publications and demonstrable tools applicable to real-world environmental challenges.

POSSIBLE HOST ORGANISATIONS

- Telecom Paris / IMT (Van-Tam Nguyen)

A-AI.2 Adaptive AI for Multimedia: Learned Compression and Real-Time Applications

Keywords: Learned Image Compression; Multimedia AI; Efficient Transformers; Real-Time Learned Video Compression.

STATE-OF-THE-ART

With multimedia making up over 50% of internet traffic, video compression is crucial for managing high-resolution, high-framerate content. Advances in AI have significantly improved image and video compression, moving beyond traditional handcrafted codecs. Learnable codecs now replace individual tools (Hybrid Image Compression) or entire codecs (Learnable Image Compression) with AI-driven models, enhancing efficiency and performance. Key advances driving this revolution include deep convolutional autoencoder architectures (CNNs) trained with VAEs and GANs, achieving compression efficiency comparable to standardized codecs. Vision Transformers (ViTs) leverage attention mechanisms to capture long-range dependencies, surpassing traditional codecs in compression ratios. Efficient Transformers like Linformer and LongFormer optimize attention complexity for faster processing in resource-constrained environments. Pruning and quantization techniques enable deployment on fixed-point embedded devices without compromising efficiency. Knowledge distillation transfers expertise from larger models to smaller ones, adapting compression performance to edge devices with limited resources.

SCIENTIFIC CHALLENGES

Despite the success of learnable codecs, research focuses mainly on improving the compression efficiency at the expense of ever-increasing complexity. Adapting learnable codecs to resource-constrained environments where computational power and memory are limited, such as edge devices, IoT systems, and mobile applications is a major problem due to the following challenges: **1) Real-time compression:** encoding or decoding HD video at 25 frames per second or above at HD resolution on smartphones or battery-operated devices is a challenging task because of the tight computational and energy budgets they allow for. This is definitely the one top challenge that needs to be tackled before learnable coding gets mainstream acceptance in the multimedia domain. **2) Bitrate adaptation:** the inherent unreliability of packet switching networks and wireless channels in particular implies that the bandwidth available to the client flutters widely over time. While traditional codecs can be easily operated under a rate control constraint, most learnable codecs attain only the rate-distortion trade-off learned at training time. **3) Domain adaptation:** while learnable codecs perform well on contents similar to those used for training, inference on dissimilar contents still result in subpar compression efficiency if not hallucination. This issue that does not affect traditional codecs currently prevents learnable image compression to be applied to specialized or sensitive domains such as screen contents or medical imaging. **4) Adapting Models Dynamically:** Developing learnable pruning mask techniques to adapt a pre-trained model on a different downstream task. **5) Adaptability:** Designing pruning algorithms that consider specific target hardware constraints.

RESEARCH ACTIVITIES

Given the above challenges, a number of different activities are proposed to address them: **1) Real-Time video compression** is the main challenge and so most of the research activities will be devoted to this problem. For hybrid coding schemes, it is proposed to explore combinations of selected AI-enabled coding tools including intra prediction, motion compensation and postprocessing. For learnable coding schemes, efficient transformer-based architectures shall be devised by resorting to the MAMBA framework and leveraging its inherent scalability over parallel hardware. Cross-domain knowledge distillation may enable selective streamlining of some model part, target different applications, e.g. user-generated or studio-generated contents. Layer pruning in place of neuron pruning will also contribute to reduce any model depth and hence its latency. **2) Bitrate adaptation**: can be achieved either via scalable coding schemes where the video is encoded at progressively better-quality layers or by adapting a baseline encoder through adapter modules. However, the implications in terms of extra complexity are not clear, especially in relationship with the above real time challenge. **3) Development of Adaptive Pruning Algorithms:** Design and implement pruning algorithms for efficiently adapting a pre-trained Learnable Image Compression (LIC) model to new rate-distortion trade-offs. Design a parameterization technique to control the level of pruning dynamically, allowing to adjust the model complexity based on specific compression scenarios or hardware constraints. **4) Unified Pruning Framework for Diverse Architettures:** Design a general pruning mask-based technique that can be applied across different LIC architectures, including transformer-based models, convolutional neural networks (CNNs), and hybrid architectures.

EXPECTED RESULTS

The above activities are expected to yield the following results: 1) Hardware benchmarking: as a main result, it is expected that real time learnable video compression can be demonstrated in a laboratory setup on mobile or edge or hardware-accelerated devices, measuring the actual trade-off between complexity and compression efficiency that can be attained. For hybrid schemes, the recently standardized H.266/VVC codec will be used as a reference. For learnable models, the sought reference will be the *Devil is in the Details* architecture with its *Window Attention Module*. **2) Bitrate adaptation:** the ability to meet different bandwidth constraints cannot be realistically demonstrated over actual wireless channels, therefore it will be simulated using pre-recorded traffic traces publicly available over a network simulator. **3) Hardware-Software Co-Design:** Optimize AI models specifically for video encoding tasks on edge hardware platforms, using techniques such as layer fusion, memory optimization, and parallel processing to improve performance. **4) Continual Learning Systems**: The compression efficiency of a learnable codec trained over natural images and adapted over computer screen contents will be evaluated both in the original and target domains; the decoded images will be evaluated for the presence of artefacts or hallucinations.

POSSIBLE HOST ORGANISATIONS

- Telecom Paris / IMT (Van-Tam Nguyen)

A-AI.3 Adaptive AI on the Edge – Innovations for Resource-Constrained Systems

Keywords: Edge AI; Adaptive AI; Resource-Constrained Systems; Energy-Efficient AI; Continual Learning; On-Device Learning; IoT

STATE-OF-THE-ART

Advancements in AI for resource-constrained systems focus on creating lightweight, efficient, and adaptable models that deliver high performance while operating within strict power, memory, and computational limits. Techniques such as quantization, pruning, and knowledge distillation have been instrumental in reducing model size and complexity, enabling deployment on edge devices. Dynamic neural networks, which adapt their structure and computation based on input or system constraints, are also emerging as a promising trend. In hardware-software co-design, innovations like layer merging, efficient memory caching, and specialized accelerators (like FPGAs and DPUs) optimize performance and energy efficiency. Advances in TinyML are pushing the boundaries of what can be achieved on microcontrollers and other ultra-low-power devices. On-device continual learning methods allow AI systems to adapt to new data in real-time without retraining from scratch, enabling personalized and context-aware applications. However, scientific challenges remain. Ensuring robustness to noisy inputs and environmental variability, achieving scalability across diverse platforms, and maintaining high accuracy while minimizing resource usage are critical barriers. Future trends point toward **real-time decision-making**, **self-optimizing models**, and **sustainable AI** practices that minimize environmental impact, with models that not only **adapt to dynamic environments** but also **optimize their performance based on resource availability**, creating new opportunities for breakthroughs real-world in edge AI deployment.

SCIENTIFIC CHALLENGES

Balancing Efficiency and Accuracy: Maintaining high model performance while optimizing for limited computation, memory, and energy resources.
Adaptability: Enabling models to dynamically adjust to changing hardware constraints and application requirements without significant performance loss.
Continual Learning: Developing on-device learning methods that prevent catastrophic forgetting and adapt to new data efficiently.
Robustness: Ensuring reliable performance under noisy, incomplete, or adversarial inputs common in edge environments.
Scalability: Designing AI solutions that scale effectively across diverse edge hardware platforms, from microcontrollers to advanced accelerators.
Real-Time Decision-Making: Enabling AI systems to make **real-time decisions** and adapt quickly to new inputs or changing environmental conditions. This includes minimizing inference latency, particularly for tasks such as autonomous navigation or robotics, where low-latency processing is critical.
Self-Optimization: Developing models that can autonomously adjust their architecture or hyperparameters based on changing workloads, data distributions, or available resources, enabling models to perform optimally without human intervention.

RESEARCH ACTIVITIES

Algorithmic Innovations: Develop lightweight AI models using techniques like pruning, quantization, and knowledge distillation; Create dynamic models and policies that can adapt their structure or computation based on resource availability or input complexity.
Continual Learning Frameworks: Design efficient on-device learning algorithms to enable models to adapt to new data while avoiding known problems like catastrophic forgetting; Implement memory-efficient methods for storing and updating knowledge during continual learning tasks on the edge. **Real-Time Adaptation Mechanisms:** Research low-latency inference methods to enable AI systems to respond in real-time; Create adaptive resource allocation algorithms to optimize energy and computational resources dynamically. **Robustness Enhancement:** Study adversarial training techniques to improve model resilience against noisy, incomplete and unlabelled inputs; Develop uncertainty policies and methods to ensure reliable decision-making under variable conditions. **Hardware-Software Co-Design:** Optimize AI models for specific edge hardware platforms, leveraging techniques like layer fusion, memory caching, and pipeline parallelism; Collaborate with hardware developers to design accelerators (e.g., FPGAs, DPUs) tailored for AI workloads. **Evaluation and Benchmarking:** Develop standardized benchmarks and metrics for assessing efficiency, robustness, and scalability of AI models in resource-constrained environments; Test AI models in real-world scenarios, such as IoT applications, robotics, and autonomous systems, to validate their effectiveness.

EXPECTED RESULTS

Efficient AI Models: Lightweight models optimized for edge devices, reducing computation and memory usage without sacrificing accuracy.
Adaptive Systems: AI systems that dynamically and/or autonomously adjust to real-time hardware and environmental constraints.
On-Device Learning: Robust continual learning frameworks for reliable real-time adaptation.
Robust AI: Models resilient to noisy or incomplete inputs, ensuring reliable performance in real-world scenarios.
Hardware-Optimized Solutions: AI integrated with edge hardware (FPGAs, DPUs, MCUs) for energy efficiency and enhanced performance.
Scalable and Interoperable Solutions: Cross-platform AI models deployable on a wide range of edge devices.
Sustainability and Real-Time AI: Low-latency, energy-efficient AI suitable for critical applications.
Comprehensive Benchmarks: Standardized benchmarks and real-world demonstrations in areas like IoT, robotics, and healthcare.

POSSIBLE HOST ORGANISATIONS

- Telecom Paris / IMT (Van-Tam Nguyen and Mazouz Alaa)

A-AI.4 LLM on the Edge

Keywords: LLM; Computation-Efficient; Memory-Efficient; Fine-Tuning; Parameter-Efficient Tuning.

STATE-OF-THE-ART

Large Language Models (LLMs) have recently demonstrated outstanding performance in a wide range of applications. This success has led to a growing demand for efficient tuning techniques to enable LLMs to adapt continuously and privately to specific tasks. However, the substantial size of these models poses significant challenges for achieving on-device adaptation, particularly on resource-constrained edge devices such as edge GPUs and smartphones.
The challenges are twofold:

- **Excessive computational overhead** during the forward and backward passes of LLMs.
- **High memory demands** for storing massive model weights and activations during the fine-tuning process.

Recent studies have shown that LLM fine-tuning typically requires state-of-the-art GPUs with large memory capacities (e.g. 40 GB or 80 GB) and can take more than a full day of computation per GPU. Even the latest efficient fine-tuning methods struggle to accommodate relatively smaller LLMs, such as LLaMA-7B, on edge devices.

SCIENTIFIC CHALLENGES

Several efforts have been made to tackle these challenges, but they have limitations:

- **Reducing computational load** often involves compressing LLMs to decrease their size. However, effectively balancing redundancy reduction with the need to preserve adaptability remains an open problem.
- **Mitigating memory overhead** usually focuses on shortening the backpropagation depth. Unfortunately, this approach limits updates to only a subset of LLM blocks, which restricts the achievable performance.

RESEARCH ACTIVITIES

Parameter-Efficient Tuning:

- Fine-tuning LLMs to new task using a limited number of trainable parameters (learnable adapters)

Memory-Efficient Tuning:

- Reduce the back-propagation depth or compression of the gradient and/or activation

Compressing then Tuning:

- Compress the LLM backbone before fine-tuning to reduce the computation and data movement overheads

EXPECTED RESULTS

- A comprehensive framework designed to address both memory and computational challenges in LLM adaptation, enabling efficient and effective deployment of LLMs on edge devices with constrained memory and computational capabilities.
- New algorithms to reduce the computation overhead.
- New algorithms to reduce the memory overhead.
- Publications in top-tier A/A* conferences or journals.

POSSIBLE HOST ORGANISATIONS

- Telecom Paris / IMT (Van-Tam Nguyen and Mazouz Alaa)

Pillar Adaptive-AI

A-AI.5 Adaptive AI-Powered Digital Twin for Innovating Healthcare Security and Resilience

Keywords: Healthcare; Adaptive AI; Digital Twins; Cybersecurity; Resilience.

STATE-OF-THE-ART

Advancements in technology and digital transformation have led to the increasing digitalization and interconnection of healthcare systems, exposing them to emerging cyber risks that threaten their resilience and continuity. Addressing these risks requires a comprehensive approach that incorporates industry insights, academic research, and government policies, ensuring robust security frameworks. Adaptive AI is playing a crucial role in healthcare cybersecurity by leveraging machine learning to continuously evolve, identify, respond to, and anticipate vulnerabilities in real time. These AI-driven systems enhance the protection of sensitive patient data, automate threat detection, and dynamically update security protocols to counter emerging risks. Digital Twins (DTs) are transforming cybersecurity in IoT-based healthcare by bridging the physical and virtual worlds through real-time data integration. These virtual models continuously monitor vulnerabilities, assess security threats, and proactively mitigate risks without impacting physical operations. DTs enhance cyber resilience by improving attack prediction, vulnerability detection, data security, and privacy. While DTs have demonstrated strong potential, real-time Adaptive AI-powered DT-based approaches are needed to further strengthen cybersecurity in healthcare applications. The integration of IoT in healthcare enables remote patient monitoring, improves service efficiency, enhances decision-making, and increases safety for patients, healthcare professionals, and staff.

SCIENTIFIC CHALLENGES

Refine Adaptive AI Mechanisms within Digital Twins for securing real-time healthcare systems: Enhance continuous learning capabilities and autonomous evolution in real-time, enabling healthcare infrastructures and services to adapt to changing circumstances, user behaviour, and dynamic environments. **Adaptive cyber resilience in healthcare:** Improve the resilience of healthcare infrastructures and services enabling them to orchestrate adaptive defences that proactively respond to anticipated cyber-attacks. **Automated Cybersecurity:** Develop models for providing continuous overview of vulnerabilities, threat landscape, attack space, and mitigate them before they happen, also for analysing complex behaviours, system architectures, and interdependencies of their components and processes. **Adaptive Anomaly Detection:** Enable dynamic closed-loop mechanisms within a virtual environment using complex behaviours to address varying dynamic cybersecurity threats and anomalies in changing environments, based on adaptive AI. **Enhancing situational awareness for emergency situations:** Ensure adaptive Digital Twins for critical IT-OT assets, including emergency vehicles, personnel, infrastructure, and industrial systems, utilize AI and machine learning techniques to simulate and visualize emergency scenarios, enhance situational awareness, and refine response strategies.

RESEARCH ACTIVITIES

Develop Adaptive AI-powered DTs: Design adaptive models (RNN, GANs and CNNs) in DT to predict real-time security threats**;** Integrate developed AI models with healthcare scenarios, i.e., elderly people in smart homes; Analysis the performance metrics for ensuring effectiveness, reliability and adaptivity**. Build DT-based architecture:** Develop a systematic way of preventing cyber threats within IoT-based healthcare systems (physically) by allowing simulation and experimentation in DT (virtually) using the developed architecture; Construct a robust framework to gather and merge data from diverse sources in real-time. **Establish dynamic threat Intelligence and prediction:** Build a sophisticated analysis platform that dynamically processes and examines data to detect potential threats; Utilize advanced algorithms and machine learning techniques incorporated with DT to analyse data trends and behavioural patterns continuously**. Develop Adaptive Learning Algorithms:** Develop advanced learning algorithms in DT that maintain high performance and accuracy as data volumes increase without forgetting learned knowledge; Adjust algorithms and DT to automatically adapt to data pattern changes, making them suitable for dynamic environments. **Develop Incident Management System:** Develop and deploy adaptive AI algorithms that analyze real-time data, predict potential incidents, assess their severity, and optimize resource allocation to prioritize emergency responses tailored to the specific needs of IT-OT environments; Integrate AI and DT with IT-OT monitoring systems to enhance incident detection, automate prioritization, and recommend optimal resource deployment strategies.

EXPECTED RESULTS

Robust set of scalable adaptive learning algorithms: ready for deployment in scenarios that require real-time data analysis and decision-making, thereby improving system responsiveness and efficiency in applications like predictive maintenance, fraud detection, and personalized services. **DT based Solution for Healthcare:** leveraging a solution to simulate different processes of physical healthcare systems dynamically adapt to environmental changes. **Cybersecurity activities module using DT in healthcare**: developing strategies to prevent the cyber threats before they happen and update the physical world about potential threats. **Security and Reliability of Testing and Validation Measures:** real-time data streaming tests to assess the generalizability and reliability of the adaptive AI models and DT technology, and a detailed report analysing the results from the validation tests, highlighting improvement in the adaptive AI models, accompanied by recommendations for future research and development. **Analysis of Scalability of DT systems** across various healthcare domains, a versatile simulated testing environment to rigorously test and refine adaptive AI algorithms under dynamic conditions., and publication of annual research findings detailing progress, breakthroughs, and strategic insights in adaptive AI research.

POSSIBLE HOST ORGANISATIONS

- NRS (Sandeep Pirbhulal and Habtamu Abie)

A-AI.6 Adaptive AI for Generalizable and Multimodal Semantic Reasoning

Keywords: Multimodal Representation; Context-Aware Learning; Explainable AI; Semantic Understanding.

STATE-OF-THE-ART

Recent advancements in image generation and editing have highlighted the limitations of relying solely on text as a conditioning modality, spurring exploration of diverse input types to better capture user intent. Sketch-based methods, for instance, offer an intuitive alternative, yet current approaches like ControlNet depend on detailed edge maps or skilled drawings. These methods struggle with abstract shapes (e.g., two arcs for a bird or a stick figure for a person) and are highly sensitive to distortions in amateur sketches, making them inaccessible to most users. Moreover, they demand significant computational resources, resulting in inefficiencies when adapting to specific concepts or user inputs in real-time. Addressing these challenges requires advancing the semantic understanding of diverse data modalities such as images, sketches, text, and audio. Models like CLIP and Vision Transformers (ViTs) have made strides in learning shared embeddings for vision and language, enabling tasks such as zero-shot learning and cross-modal retrieval. Recent approaches, including visual prompt learning, attention refinement, and hierarchical training strategies, enhance semantic disentanglement and domain transferability while reducing reliance on resource-intensive annotations. Nonetheless, several key challenges remain: (i) Generalizing to underexplored modalities like freehand sketches, diagrams, and 3D visualizations; (ii) Handling noisy, ambiguous, or sparse inputs without compromising interpretability; (iii) Designing efficient, lightweight architectures suitable for edge or resource-constrained environments. The convergence of adaptive AI and advanced semantic understanding is vital for creating systems that are robust, scalable, and capable of real-time adaptation, empowering broader accessibility and context-aware functionality.

SCIENTIFIC CHALLENGES

To make AI more accessible, it is essential to eliminate the need for large computational resources and long wait times that typically plague image and video generation models. A key challenge lies in improving the controllability of these models, particularly in understanding abstract representations and handling deformations in user inputs, such as sketches. By fostering creativity in image generation models, we can make AI more intuitive and responsive to diverse user needs. Semantic understanding in AI faces significant hurdles in achieving generalization, adaptability, and efficiency across diverse modalities, including visual, textual, and abstract inputs like sketches and diagrams. Many current models struggle to seamlessly operate across these different input types while maintaining robustness in low-resource or noisy environments. AI systems must also be capable of dynamic adaptation to new domains or tasks, incorporating contextual signals, such as temporal dependencies or user feedback, to improve their flexibility. Scaling these systems for deployment on resource-constrained devices, such as edge hardware, further complicates the design, requiring lightweight architectures that balance computational power with accuracy. Finally, aligning AI-generated outputs with human-level understanding, especially in ambiguous or abstract scenarios, is crucial for enhancing trust, usability, and explainability.

RESEARCH ACTIVITIES

To address these challenges, research must focus on designing unified semantic representation models that generalize across multiple modalities, such as images, sketches, and videos, while leveraging weak supervision and noisy labels to reduce annotation dependence. Context-aware adaptation mechanisms are essential for refining AI systems dynamically using hierarchical and cross-attention strategies. Additionally, lightweight model designs, such as pruning, quantization, and dynamic neural networks, must be explored to ensure efficient on-device deployment. Human-centric approaches, including human-in-the-loop experiments, will play a vital role in aligning model outputs with user expectations and enhancing explainability. To evaluate progress, benchmarks that encompass multimodal and abstract data, as well as metrics for adaptability and resource efficiency, must be established to ensure relevance in real-world scenarios.

EXPECTED RESULTS

To address these challenges, research must focus on developing unified semantic representation models that generalize across multiple modalities like images, sketches, and videos, while reducing reliance on annotations through weak supervision and noisy labels. Context-aware adaptation mechanisms, such as hierarchical and cross-attention strategies, are vital for refining AI systems dynamically. Furthermore, exploring lightweight model designs—such as pruning, quantization, and dynamic neural networks—is necessary to ensure efficient on-device deployment. Human-centric approaches, including human-in-the-loop experiments, will help align model outputs with user expectations, improving explainability. Finally, establishing benchmarks and metrics that assess adaptability and resource efficiency in real-world scenarios will be essential for driving progress in these areas.

POSSIBLE HOST ORGANISATIONS

- Telecom Paris / IMT (Van-Tam Nguyen and Mazouz Alaa)

A-AI.7 Zero-shot large-scale biomedical entity matching and linking

Keywords: Entity Linking; LLMs; Zero-Shot Domain Adaptation; Domain Transfer; Knowledge-Based Transfer.

STATE-OF-THE-ART

The advent of large language models has allowed the development of supervised entity-linking approaches that far surpass knowledge-based matching. However, these approaches require large, annotated corpora for training and a specific corpus is only linked to specific ontologies. In the general domain, the largest corpora are typically based on Wikipedia and contain Wikidata or DBPédia entities. While for general purpose knowledge extraction, this might not be a significant problem due to a rich coverage of Wikidata or DBPédia, but for domains with very specific ontologies, covering very specific areas, zero-shot or dictionary-based approaches that transfer to any ontology are required. This is the case is the clinical and biomedical domains, where semantic annotation tools on platforms such as Bioportal rely on dictionary-based approaches, which do not leverage the latest advances in deep-learning based knowledge extraction. The latest zero-shot approaches, no longer require occurrence tables or other information besides a dictionary of concepts with associated definitions. Definitions can be used to fine-tune LLMs and learn concept representations agnostic from any given knowledge base, that allow the models to transfer to unseen ontologies. Even then, such approaches still rely on an entity matching or recognition system to identify mentions to be linked. If we want to build a true end-to-end zero-shot entity linking system independent of NER and of annotated corpora, then we must look at the zero-shot NER literature, which offers some more general transfer approaches that are able to learn span representations that enable models to identify most likely spans without explicit supervised training (e.g. by leveraging silver data from string matching approaches). An incremental sliding-window approach could allow to first identify groups of tokens corresponding to most likely mentions (metric learning approach), and then using concept representations (based on definitions, knowledge embeddings or both) and learn-to-rank losses identify the most likely linked concepts.

SCIENTIFIC CHALLENGES

- Effectively exploit silver-data for semi-supervised fine-tuning by devising an appropriate loss function for metric learning in order to remove the dependency on a NER system.
- Identify the best knowledge infusion and concept representation learning strategy to implement the learning-to-rank-approach.
- Ensure computational efficiency in order to cater to large scale annotation services (e.g. replacing the Bioportal Annotator), perhaps using approaches like adapter (e.g. LoRA).

RESEARCH ACTIVITIES

- Explore the literature from Zero-shot NER and Zero-shot EL to pin-point relevant methodological approaches.
- Build a model based on pre-trained clinical LLMs to identify likely mentions using a combination of supervised fine-tuning on large clinical corpora (e.g. MIMIC IV) and of string-matching mentions identification. The latter being more uncertain and prone to errors, the uncertainty associated should be integrated in the loss formulation.
- Extend the model with learning-to-rank approaches using both supervised fine-tuning, and knowledge infusion (textual concept embeddings, KG embeddings, etc.) for unsupervised adaptation to new ontologies using only unannotated text.
- Evaluate on a broad set of clinical EL tasks.

EXPECTED RESULTS

- A preprint and subsequent journal (Journal of Biomedical Semantic) or conference (*ACL) publication that reports the research undertaken and the findings.

POSSIBLE HOST ORGANISATIONS

- Telecom Paris/IMT (Andon Tchechmedjiev)

A-AI.8 Parameter Efficient Algorithms for Foundation models

Keywords: Parameter-Efficient Techniques LLM; VLM; Continual Learning.

STATE-OF-THE-ART

Foundation models, such as Large Language Models (LLMs) and Vision Language Models (VLMs), have revolutionized AI by scaling across diverse tasks and modalities. These models leverage extensive pre-training on large corpora, capturing versatile representations that support zero-shot and few-shot learning. However, along with computational demands, they also tend to suffer from forgetting when trained on new sequential tasks. Continual Learning, enables models to adapt incrementally to new tasks and data distributions without forgetting prior knowledge. Developing parameter-efficient algorithms for foundation models requires optimizing adaptation while preserving pre-trained knowledge, ensuring robust and efficient performance across tasks and modalities.

SCIENTIFIC CHALLENGES

Pre-trained models are trained on a huge corpus of data; however, we do need to train when adapting to a new task or distribution. Developing parameter-efficient algorithms for pre-trained models involves significant challenges, particularly in adapting to new tasks or data without losing information from prior training. Parameter-efficient fine-tuning (PEFT) techniques aim to address this by optimizing adaptation processes, but achieving a balance between efficiency and performance remains a critical hurdle. Moreover, continual learning introduces additional difficulties, such as preventing catastrophic forgetting and managing task interference, especially under resource constraints. The ultimate challenge is to design algorithms that efficiently expand knowledge over time while preserving and enhancing the performance of previously learned tasks, ensuring scalability, adaptability, and robustness in diverse real-world applications.

RESEARCH ACTIVITIES

- Comprehensive literature review on PEFT techniques– vision models or LLMs or VLMs
- Explore PEFT methods in continual learning
- Explore different techniques of efficiency – sparsity, pruning, prompting, adapters
- Explore how different modalities (vision, language) help or deter in this process
- Propose a new methodology based on insights from the evaluations, aiming to surpass current state-of-the-art techniques.

EXPECTED RESULTS

- Evaluation of different PEFT methods in supervised learning, continual learning
- A novel PEFT method for efficient learning in single or multi-modal models
- Publications in top-tier A/A* conferences or journals.

POSSIBLE HOST ORGANISATIONS

- TU/e (Shruthi Gowda and Bahram Zonooz)

A-AI.9 Robustness and Generalization in single or multi-modal models

Keywords: Robustness; Generalization; LLM; VLM; Supervised Learning; Continual Learning.

STATE-OF-THE-ART

Deep neural networks (DNNs) have achieved remarkable success across various domains; however, they still lag behind human cognitive abilities, particularly in robust learning and adaptability. Humans can seamlessly adjust to new information, whereas DNNs are susceptible to adversarial attacks—minor input perturbations that lead to significant misclassifications. With the emergence of multi-modal architectures, analysing robustness in different settings and achieving robustness comparable to human resilience remains a complex and ongoing challenge.

SCIENTIFIC CHALLENGES

While humans exhibit robust learning and adaptability, deep neural networks (DNNs) are vulnerable to adversarial attacks, easily misled by minor input alterations. These limitations highlight the need for advanced, brain-inspired approaches in AI research. The primary objective is to enhance robustness and generalization against both natural and adversarial perturbations. This involves developing novel training schemes to address the trade-off between standard and adversarial accuracy and tackling robust overfitting, ensuring the development of comprehensive AI systems capable of reliable performance in diverse and challenging scenarios. Analysing robustness across vision-only models, language models, and multi-modal models is crucial, as it provides insights into how different modalities impact overall system resilience.

RESEARCH ACTIVITIES

- Comprehensive literature review on robustness – vision models or LLMs or VLMs
- Explore and decide on the setting – supervised learning, continual learning
- Evaluate the robustness of single and multi-modal models, exploring how different modalities may enhance or hinder each other's performance.
- Propose a new methodology based on insights from the evaluations, aiming to surpass current state-of-the-art techniques.

EXPECTED RESULTS

- Comprehensive analysis on robustness of single or multimodal models on different settings
- New algorithms to improve robustness
- Publications in top-tier A/A* conferences or journals.

POSSIBLE HOST ORGANISATIONS

- TU/e (Shruthi Gowda and Bahram Zonooz)

Pillar Human-Centric AI

HC-AI.1 Interpretability and uncertainty in predictive models

Keywords: Explainable AI; Uncertainty Quantification; Conformal Prediction; Trustworthy AI.

STATE-OF-THE-ART

AI models that score high on both accuracy and interpretability are in demand. Traditional post-hoc explanation methods often suffer from inconsistencies and lack of fidelity to the underlying models, which has led to increased emphasis on developing models that are more transparent and interpretable by design. Simultaneously, accurate uncertainty quantification in predictive models is essential not only for informed decision-making, but also for instilling trust in automated systems and ensuring their safety. Failure to properly account for uncertainty can lead decision-makers to inappropriately trust potentially incorrect outputs, undermining confidence even in accurate predictions.

While current approaches often treat interpretability and uncertainty quantification objectives in isolation, there is interesting recent work on bringing the two domains closer together. For instance, it is possible to use the framework of Conformal Prediction (CP) – a robust, model-agnostic approach to uncertainty quantification – to quantify the uncertainty associated with explanations. Besides such uncertainty-aware explanations, there have also been recent advances in the design of intrinsically interpretable models with guaranteed accuracy of explanations, employing sparse representations to evade problems with first learning a complex model and then finding post-hoc explanations.

SCIENTIFIC CHALLENGES

We invite research project proposals that tackle some of the following challenges:

- Developing new models that integrate interpretability and uncertainty quantification, enabling human-centric, trustworthy predictions and explanations.
- Exploring and extending existing methodologies like InterpretCC for various prediction tasks with relevance for telecom applications, for instance multivariate, irregular timeseries forecasting.
- Creating metrics and evaluating real-world applications to assess how interpretability and/or uncertainty information influence users' trust and the usability of critical systems.

RESEARCH ACTIVITIES

- **Literature review:** Study existing methods combining interpretability and uncertainty quantification.
- **Method development and benchmarking:** Develop and evaluate methods suitable for various telecom-relevant prediction tasks, benchmarking them on synthetic and open-source datasets – for examples, see Ref. [7].
- **Real-world evaluation:** Testing selected interpretability and uncertainty quantification methods on users, measuring their impact on tasks like forecasting and anomaly detection.

EXPECTED RESULTS

We expect the exchange to provide valuable contributions to:

- **Refined methods:** Improving ML algorithms for prediction tasks, with integrated interpretability and uncertainty features.
- **Practical applications:** Demonstrating reliability in real-world use cases such as time series forecasting and anomaly detection.
- **Human-centric trust:** Enhancing user trust in AI systems through accurate uncertainty estimation and interpretable methodologies.
- **Dissemination:** Publishing findings in a high-impact conference or journal to engage the research community and drive advancements.

POSSIBLE HOST ORGANISATIONS

- Telenor Research & Innovation

Pillar Human-Centric AI

HC-AI.2 Improving transparency and explainability of web-based AI systems through semi-structured natural language descriptions

Keywords: Natural Language; Scenarios; Natural Language Processing; Transparency; Trustworthy AI.

STATE-OF-THE-ART

Web systems feature complex architectures integrating multiple services and increasingly AI components. For end-users, it is thus hard to see what is happening behind the frontend. This particularly impacts transparency and explainability of these systems, which are key factors for trustworthiness. Existing architectural modeling approaches such as UML, SysML, or functional block diagrams do not address these aspects well. Due to their complex visual semantics, they require technical expertise to be understood, and they do not sufficiently represent information that is relevant for the trustworthiness of the distributed AI system. Thus, there is a need for a novel descriptive approach capable of representing relevant factors of trustworthiness for AI web systems.

SCIENTIFIC CHALLENGES

Elicitation and representation of problem and solution domain knowledge relevant to the trustworthiness of complex web systems with AI components need to be supported by techniques suitable for various technical and non-technical stakeholders. This is important to leverage information exchange between requirement engineers, AI experts, system architects, and ultimately end-users of the systems. Techniques based on natural language (NL) have long been successfully used in such scenarios. However, it remains unexplored whether semi-structured NL artifacts such as scenarios are applicable for the trustworthiness of AI in complex web system architectures and, if yes, which modifications are necessary. Furthermore, the integration of NL-based representations with suitable visual architecture modeling approaches needs to be investigated. Automation or partial automation by means of NLP/AI techniques promises greater applicability in settings with effort and resource constraints.

RESEARCH ACTIVITIES

We invite researchers to collaborate on several or ideally all of the research activities described in the following. A survey of existing NL-based approaches and their applications in the context of trustworthy AI and architectural modeling needs to be conducted. Based on inputs from the ENFIELD taxonomy of trustworthy AI and exemplary use cases, suitable semi-structured NL representations need to be analysed and adapted. The resulting models shall be co-designed with visual modeling approaches and their integration demonstrated. Lastly, NLP and AI algorithms to automate the proposed method need to be developed and demonstrated.

EXPECTED RESULTS

We expect the exchange to provide valuable contributions to the long-term goal of designing a method for architectural trust modeling in complex distributed AI systems. The method will facilitate the creation of distributed “trustworthy by design” AI systems by enabling system architects to document and analyse trust in their architectural system blueprints. The outputs of this challenge will be a method for NL-based representation of AI trustworthiness information in web-based systems, a prototype demonstrating the method through integration with a visual modeling editor, and techniques for automating the method. The exchange also aims to foster knowledge transfer and networking with other groups with a demonstrated competency in requirements engineering, software engineering and natural language processing.

POSSIBLE HOST ORGANISATIONS

- TUC (Sebastian Heil)

HC-AI.3 Explainable AI for Multimodal and Sequential Data Analysis in Physical and Chemical Processes.

Keywords: Explainable Artificial Intelligence (XAI); Multimodal Data; Sequential Data; Machine Learning; Interpretability; Materials Science.

STATE-OF-THE-ART

The rapid advancements in machine learning (ML) have transformed data-driven approaches to analysing complex systems, particularly in domains requiring the integration of multimodal or sequential data. Despite these achievements, the interpretability of ML models remains a critical barrier. While explainable AI (XAI) methods have been applied extensively to image and tabular data, their application to temporal and multimodal datasets presents unique challenges due to the complex interdependencies and non-linear relationships inherent in such data. In fields such as material sciences, electrochemical, and biomedical research, techniques like electrochemical impedance spectroscopy (EIS) or other multimodal analyses often involve intricate sequential data. Current state-of-the-art ML methods can predict outcomes or classify patterns from such datasets with high accuracy. However, they often fail to elucidate the underlying relationships between the input features and the prediction outputs, limiting their utility for guiding scientific discovery or decision-making. Addressing this gap requires novel approaches that integrate advanced XAI techniques with domain-specific knowledge to provide actionable insights.

SCIENTIFIC CHALLENGES

The scientific challenges in explainable AI (XAI) for multimodal and sequential data analysis lie in four key areas. First, there is a need to design machine learning models that not only achieve high predictive accuracy but also provide interpretable insights into the underlying physical or chemical processes in complex systems. Second, the integration of data from multiple modalities or time-series measurements creates high-dimensional spaces, where extracting meaningful and actionable explanations remains both computationally challenging and conceptually underexplored. Third, XAI tools must be tailored to generate insights that are specifically relevant to domain-specific applications, such as optimizing materials or designing advanced devices, while avoiding pitfalls like overfitting or introducing biases inherent to the domain. Finally, these models must be computationally efficient and scalable, capable of handling large datasets and adapting to real-time applications without compromising explainability.

RESEARCH ACTIVITIES

To address these challenges, the proposed research will encompass the following activities:

- **Development of Explainable Models:** Design and train ML models that inherently incorporate mechanisms for feature attribution, counterfactual analysis, and other XAI techniques, specifically adapted for multimodal and sequential data coming from physical or chemical processes.
- **Simulation and Dataset Generation:** Create synthetic datasets from domain-relevant models to establish ground truth for explainability experiments. Complement this with real-world datasets to validate generalizability.
- **XAI Validation Through Experts:** Involve domain experts in the evaluation of XAI frameworks to assess the relevance, clarity, and utility of the generated explanations. Expert validation will ensure that the insights provided by the models align with scientific principles and guide further experimentation or design.

EXPECTED RESULTS

The expected outcomes of this research include the development of XAI algorithms specifically designed for sequential and multimodal data from physical or chemical processes, integrating domain knowledge to disentangle complex, interdependent features. These algorithms will empower researchers with actionable scientific insights, providing interpretable outputs that facilitate the understanding and optimization of underlying processes in applications such as material sciences. Additionally, the project aims to produce open-source frameworks and benchmarks, offering comprehensive datasets and reusable tools to support further advancements in XAI for sequence and multimodal data analysis. Validated by domain experts, these XAI tools will meet the scientific rigor and practical requirements of specific fields, ensuring their applicability in real-world optimization and design tasks. Finally, the results should be submitted to at least one top-tier journal or conference in machine learning, artificial intelligence, or material sciences.

POSSIBLE HOST ORGANISATIONS

- TU/e (Isel Grau)

T-AI.1 Security and Robustness of AI systems
Keywords: AI security/robustness; trustworthiness; LLMs; adversarial machine learning; verifiability; uncertainty quantification.
STATE-OF-THE-ART
The rapid shift of the form of AI systems has introduced several challenges related to the security and robustness of such systems and consequently to their trustworthiness. The continuous growth and development of new techniques and methodologies in the AI field makes security of related systems even more difficult to achieve as the attack vectors are expanded with every new advancement in the field. Additionally, ensuring robust performance and accurate uncertainty quantification in AI models is crucial for maintaining trust in automated systems. The state of the art refers mainly to research on adversarial machine learning, relevant mitigation measures in the traditional machine learning field, and uncertainty estimation techniques like conformal prediction. It is of utmost importance to conduct research on the security and robustness of current state-of-the-art AI systems such as LLMs and indicate new possible attacks and/or provide relevant solutions.
SCIENTIFIC CHALLENGES
The main research challenges relate to: (1) adversarial machine learning attacks, (2) adversarial machine learning detection, (3) adversarial machine learning defences (4) LLMs related attacks such as prompt hacking or adversarial attacks, (5) LLMs defences, (6) AI fairness, (7) AI security by design approaches, (8) monitoring and measuring AI systems security (9) means for verifiable training and/or inference in AI systems, (10) uncertainty quantification in AI systems (e.g., timeseries forecasting) to improve robustness, and (11) calibration techniques to address biases and enhance model reliability.
RESEARCH ACTIVITIES
Research on any security or robustness aspect of traditional or modern AI systems. The activities expected will be related to all or some of the following: (1) identification of an interesting topic in the relevant research area, (2) literature survey for the selected topic, (3) proposal for a novel approach for that topic, (4) development of a proof of concept for the proposed approach, (5) comparison of the proposed approach with similar approaches in literature, (6) reasoning about the significance of the approach, (7) preparation of a relevant paper and (8) submission of the paper to a related venue.
EXPECTED RESULTS
The ENFIELD project will leverage novel scientific results to increase the trustworthiness of AI. By leveraging the results from this topic directions and guidelines towards the development of a trustworthy AI framework for EU will be facilitated. In addition to that, the involved partners and research will collaborate, exchange knowledge, and expertise to further develop their research activities and future collaborations. It is required to produce at least one scientific publication out of this collaboration
POSSIBLE HOST ORGANISATIONS
• NTNU (Georgios Spathoulas) • Telenor Research & Innovation (Jeriek Van den Abeele and Claudia Battistin)

T-AI.2 Privacy and Compliance of AI systems

Keywords: AI Privacy; AI Compliance; Sensitive Data; Homomorphic Encryption; Federated Learning.

STATE-OF-THE-ART

AI systems process a vast amount of data (personal/sensitive). and they are strongly required to conform to the relevant regulations (GDPR). On top of that, it is important to advance state of the art with respect to technical measures that can facilitate privacy protection of data and AI models. The recent advancements of LLMs have brought new issues with regards to training data availability, consent of users to have their personal data included in training datasets and access control to closed access AI models the use of which is offered under an AI as a Service scheme.

SCIENTIFIC CHALLENGES

As AI systems become prominent in our lives it is important to deal with privacy requirements and enable regulations that can be practically applied to real world systems. The main research challenges relate to: (1) identification of privacy leakage in AI systems, (2) methodologies to make AI systems more privacy friendly, (3) use of cryptographic techniques (homomorphic encryption, zero knowledge proofs) to enhance privacy, (4) use of federated learning approaches to increase privacy in distributed setups, (5) regulatory framework for AI systems, (6) tooling to monitor/prove regulatory compliance, (7) users perspective on trustworthy AI and relevant privacy issues and (8) approaches to increase users' awareness.

RESEARCH ACTIVITIES

Research on any privacy and regulatory aspect of traditional or modern AI systems. The activities expected will be related to all or some of the following: (1) identification of an interesting topic in the relevant research area, (2) literature survey for the selected topic, (3) proposal for a novel approach for that topic, (4) development of a proof of concept for the proposed approach, (5) comparison of the proposed approach with similar approaches in literature, (6) reasoning about the significance of the approach, (7) preparation of a relevant paper and (8) submission of the paper to a related venue.

EXPECTED RESULTS

ENFIELD project expects that the research work towards the specific challenge will provide advancements to achieving privacy preservation (to the extent that this is feasible) in AI systems. Alternatively, collaboration can enhance regulatory frameworks that are coming up globally and provide tooling relevant to their practical application. The researcher may work towards novel AI workflows and approaches that will facilitate the development of trustworthy AI systems that are compliant with such frameworks. The project expects at least one scientific publication in one of the mentioned areas as the outcome of the exchange and the formation of a collaboration between the visiting researcher and the hosting institution that can be extended even after the research visit.

- NTNU (Georgios Spathoulas)
- TUC
- Telenor Research & Innovation (Jeriek Van den Abeele)

Pillar Trustworthy AI

T-AI.3 Trustworthy ML based scheduling for the energy domain

Keywords: Trustworthiness of AI; AI Robustness; Adversarial Machine Learning; Uncertainty Quantification; Energy Usage Monitoring; Secured Digital Twin; AI Trust Modelling.

STATE-OF-THE-ART

The production scheduling system entwined with its digital twin efficiently allocates resources, machines, and workers to a set of tasks or jobs, while minimizing time, cost, or resource usage; while providing a simulation-based digital environment where multiple AI trustworthiness experiments can be carried out. The growing concern regarding the dependability of ML algorithms, which cannot be entirely trusted due to their fragile nature leads us to a dire need for systematic analysis of adversarial settings in real-world scenarios. The production scheduling will involve introducing subtle manipulations or disruptions in a way that can negatively affect the performance of a system without being immediately obvious. Such an attack would aim to disrupt the optimality of schedules, increase costs, or cause delays while appearing as legitimate inputs or perturbations which causes a ripple effect in the overall production schedule. Hence establishing a more secured production line by highlighting the trustworthiness of AI by the dint of adversarial ML algorithms is the need of the hour.

SCIENTIFIC CHALLENGES

Creation of an optimised production scheduling based on data ingestion pipelines of orders, machine capacity and energy behaviour of machines (linea robot, roland 307 and imbustatrice) by considering the daily data ingestion and historical data sets. Running an adversarial ML attack on the digital twin model of the production line provided by Maggioli subtly manipulates input parameters in a production line such as job durations, machine availability, order quantity, order priority or modify the environment (e.g., unexpected failures, delays) to disrupt the optimal schedule. Purposely design to cause a model to make a mistake in its predictions despite resembling as a valid input to humans.

The digital twin environment will facilitate real-time data handling and develop a data management component to run energy aware analytics engine by training and executing adversarial ML Models. Model energy meters and operational data sources logbook will also serve as metadata for data granularities. Model 'what-if' scenarios and root causes mapping to understand impact of energy and uncertainty quantification in AI systems (e.g., timeseries forecasting) to improve AI robustness, AI trustworthiness and outputs from adversarial ML algorithms for re-calibration techniques to address biases and enhance AI model reliability.

RESEARCH ACTIVITIES

The research activities will focus on trustworthiness, security and robustness aspects of AI systems. The activities expected will be related to creation of optimised production scheduling based on adversarial ML attack on the digital twin model platform. Modelling of the machine behaviour based on calculation of optimal solutions and verify with prediction and simulation services, by running the adversarial ML algorithms.

EXPECTED RESULTS

The ENFIELD project aims to advance trustworthy, secure, and robust AI systems. Researcher(s) will develop efficient and secured production scheduling systems. Expected outcomes include at least one scientific publication and a lasting collaboration between the visiting researcher and the hosting institution, extending beyond the research visit.

POSSIBLE HOST ORGANISATIONS

- MAGGIOLI (Andrea Montefiori and Astik Samal)
- NTNU (Georgios Spathoulas)

Pillar Trustworthy AI

T-AI.4 AI in Distributed Systems

Keywords: Trust modelling; Edge intelligence; AIaaS; Software Architecture; Federated learning; Risk assessment.

STATE-OF-THE-ART

The integration of artificial intelligence is rapidly increasing and becoming more prevalent in our daily routines and the systems we interact with regularly, such as in healthcare, finance, transportation, social media, and online services. Those systems are utilizing AI to automate analysis and processing tasks and are becoming integral in decision-making processes. The underlying architecture of the systems is typically distributed, integrating AI services as system components or implementing a distributed AI system on its own. New trust-related challenges arise from the interplay of the distributed nature of the system architecture with the specific characteristics of AI components. These challenges need to be addressed in all phases of the system's lifecycle: in the architectural design of the system as well as during development, testing, maintenance and operation. Addressing the trust challenges at an early stage helps in creating a resilient architecture capable of supporting the dynamic and distributed nature of modern systems, ultimately enhancing the overall trustworthiness of the system. Existing modelling and analysis techniques for distributed systems lack a systematic consideration of trust aspects and AI-related characteristics.

SCIENTIFIC CHALLENGES

AI components are used in different parts of complex distributed or even federated systems, which raises challenges. One challenge that arises from integrating AI is that it impacts the trustworthiness of the overall system architecture. The unpredictability and opacity of AI components can harm users' trust in the system. This issue is particularly critical because AI models are often considered "black boxes", making it difficult to predict their behaviour and understand their decision-making processes. Building trustworthy AI systems requires implementing robust verification and validation techniques throughout AI training and inference phases. Ensuring the transparency, robustness, and fairness of AI systems is essential to maintaining user trust. Another challenge that arises due to integrating multiple AI components within a distributed system is that these components can interact with each other without clear verification or understanding, increasing the complexity of the entire system. These AI-to-AI interactions can be unpredictable and lack transparency, leading to potential issues in system reliability and trust. Increasing reliance of distributed systems on third-party AI services (AIaaS) poses another challenge. AI service consumers need to be able to ensure that the actual model and configuration used by the provider aligns with the intended model and configuration, which affects the system's verifiability and trustworthiness. Due to the nature of AI models, this cannot be verified on the results. Thus, the AIaaS providers can use an alternative model or manipulating the model parameters to reduce running costs and utilize less energy without the consumer's knowledge. Service Consumers currently have only limited means to verify the integrity and performance of third-party AI models. To address the above challenge of distributed AI systems, we are particularly interested but not limited to contributions in the following areas: Modelling and analysing trust on the architectural level; Users' perception of trust; Trust in Federated learning; Verifiability of AI inference in AIaaS scenarios; LLMs to assist trust and risk assessment; Explainable AI.

RESEARCH ACTIVITIES

We invite researchers to collaborate on one or more of the following research activities: the extension of a taxonomy of trust in distributed AI system architecture; the specification of a suitable visual modeling language; the development of infrastructure supporting the modeling; the design of algorithms for automatic analyses; the evaluation of trust modeling in distributed AI systems. Other research activities related to the challenges are described above.

EXPECTED RESULTS

We expect the exchange to provide valuable contributions to the long-term goal of designing a method for architectural trust modeling in complex distributed AI systems. The method will facilitate the creation of distributed "trustworthy by design" AI systems by enabling system architects to document and analyse trust in their architectural system blueprints. For researchers, the results will contribute to establishing a common vocabulary and representation of trust in distributed AI systems as a first step to consolidate the body of knowledge in this relatively young field and facilitate the communication and thus collaboration. The exchange also aims at fostering knowledge transfer and networking with other groups working in related fields such as information systems, distributed systems and software engineering.

POSSIBLE HOST ORGANISATIONS

- NTNU (Georgios Spathoulas).
- TUC (Sebastian Heil).

T-AI.5 Assessing Trustworthiness of Distributed AI Systems

Keywords: Scale Development; Construct Validity; Construct Reliability; HCI.

STATE-OF-THE-ART

AI-integrated web systems are increasingly prevalent in areas like fintech, social media, eHealth, and eCommerce, where trust and trustworthiness are crucial for long-term adoption. Security analysis focuses on secure communication and information sharing to ensure resilience. Trustworthiness is viewed as a composite non-functional property encompassing attributes like safety, security, availability, and reliability. The STRAM framework, introduced by Cho et al., uses system-level metrics and an ontological approach to assess trustworthiness. While system-centered methods are essential, a holistic approach is needed to address social and ethical impacts on both end-users and society.

SCIENTIFIC CHALLENGES

Artificial Intelligence automates tasks that demand efficient data processing and analysis in within complex web system architectures and user acceptance and adoption depend on the performance and reliability of the systems. However, high performance and reliability of AI-enhanced web systems often comes at the expense of system transparency and explainability, especially in cases where black-box machine learning models are employed as core architecture components. Since the systems are not immune to failure and errors which, depending on the application domain and use case, can potentially have devastating consequences, over-trust in these systems is important to understand and address. Ensuring that web-based AI systems are inherently trustworthy is a challenge that goes beyond technical perspectives of system architecture and cybersecurity. Measurement is the basis for systematic engineering approaches, allowing to inform design decisions and evaluate outcomes. Thus, a major challenge lies in the creation of suitable instruments that operationalize trustworthiness of web systems with AI components, considering the perspective of both developers and end-users and how interactions with these systems influence the perceived trust in them. Human-in-the-loop approaches to the development of trustworthiness instruments for AI-enhanced web systems present unique problems and require innovative thinking to provide both theoretical insight and practical solutions to this challenge. Research needs to also reflect on how trustworthiness is part of the broader framework of social reality. Another challenge for the creation of suitable measurement instruments lies in investigating to what extent trustworthiness properties from a system-centered and human-centered perspective overlap or divert. To address such fundamental challenges, the ability to apply and combine knowledge and methods from various fields of computer science (e.g. software engineering, human-computer interaction, cybersecurity, networking and AI) is critical. Solutions need to synthesize expertise in the topic of trustworthiness and how the concept relates to neighbouring notions of transparency, explainability, fairness, and accountability to ensure that they will resonate with developers and end-users. To adequately address the challenges, contributions are encouraged (but not limited) to the following areas: multidisciplinary perspectives on trust theory, analysis and modeling of trustworthiness, measurement and scale development, human-centered explainable AI, user preference research and formal methods. The nature of the proposed challenge requires significant collaboration across disciplinary boundaries, and insights from the area of HCI (combining knowledge and methods from humanities, social sciences and arts). Considering such human-centered perspective will support long-term benefits of the developed trustworthiness instrument and help to establish a joint understanding of the unique challenges of trustworthy AI in distributed systems.

RESEARCH ACTIVITIES

A successful TES collaboration to meet the proposed challenge will consist of activities related to (1) conducting a systematic review of the trustworthiness concept relevant to AI in web systems, (2) operationalizing the relevant trustworthiness conceptualization to specific architectural characteristics of trustworthy AI in web systems, (3) developing a trustworthiness instrument (e.g. scale, metric, checklist etc.) that can support developers with assessing how trustworthy the system architecture is perceived depending on specific user groups or application domains (4) evaluating the developed instrument with users as a first step of validation.

EXPECTED RESULTS

We expect the exchange to provide valuable contributions to the long-term goal of ensuring and monitoring the trustworthiness of complex distributed AI systems throughout their lifecycle. The outputs of this challenge will provide developers with a concrete trustworthiness instrument to ensure that the AI distributed systems are considering trustworthiness also from the end-user perspective, which will increase their acceptance and use. A common understanding of how to assess and design for trustworthy AI in web systems among both developers and end-users will also aid in broader ethical discussion about the need for their accountability and regulation. As such, the expected outcome also aims to strengthen theory-driven research into the development of trustworthy AI distributed systems. The exchange also aims at fostering knowledge transfer and networking with other groups with a demonstrated competency in HCI, social, ethical, and science-philosophical perspectives on Trustworthy AI.

POSSIBLE HOST ORGANISATIONS

- TUC (Sebastian Heil).

T-AI.6 Brain-to-Speech Interface: From Neural Signals to Communication Restoration

Keywords: BCI; Speech Restoration; Neural Vocoder; EEG Decoding; Deep Learning.

STATE-OF-THE-ART

Brain-computer interfaces (BCIs) for speech restoration are a new and important area of research with real-world benefits, especially for people who cannot speak due to serious conditions. Recent progress in computer algorithms, deep learning, and speech synthesis tools has made it possible to turn brain signals into speech. This gives hope for helping people communicate again. However, there are still challenges, like making this technology accurate, easy to use, and practical for hospitals or clinics. Researchers are working on improving how well brain signals can be understood, exploring new ways to process EEG data, and creating better speech models to make the reconstructed speech sound more natural and clearer.

SCIENTIFIC CHALLENGES

The main challenge is to create a reliable brain-to-speech system that can decode speech patterns from EEG data to help people with communication difficulties. This involves addressing several key areas:

- Developing better methods to clean and process noisy EEG data, focusing on the brain signals linked to speech.
- Designing and improving deep learning models, such as transformer-based systems, for accurate decoding of speech patterns.
- Integrating neural vocoders like AutoVocoder or BigVGAN to produce clear and realistic speech.

Building real-time systems that connect brain signals to synthesized speech for practical use in clinical settings.

RESEARCH ACTIVITIES

The activities will include:

- Literature review of brain-to-speech technologies and EEG signal decoding techniques.
- Preprocessing of multi-channel EEG data using denoising autoencoders, wavelet transformations, and Hilbert-Huang analysis.
- Developing deep learning models to decode speech envelopes, employing advanced architectures such as transformers and VAEs.
- Synthesizing decoded signals using neural vocoder systems to evaluate naturalness, intelligibility, and real-time performance.
- Designing experiments to benchmark models against existing solutions, followed by preparing and submitting scientific publications.

EXPECTED RESULTS

The ENFIELD project anticipates innovative solutions that advance EEG-based speech decoding and synthesis. These include:

- A scalable and clinically applicable brain-to-speech interface.
- At least one high-quality publication on decoding and synthesis techniques.

Collaboration between academic and clinical research teams to enabling knowledge exchange.

POSSIBLE HOST ORGANISATIONS

- BME (Géza Németh)

T-AI.7 Secure Voice Biometrics with Fake Voice Detection

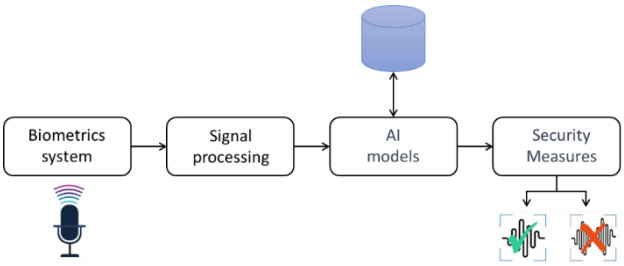
Keywords: Voice spoofing; Biometric security; Speech signal processing; Robust authentication; Acoustic analysis.

STATE-OF-THE-ART

Current voice biometric systems are at the central of biometric authentication, but they face increasing concerns related to data privacy and security. The rise of fake voice generation technologies presents a significant challenge to the integrity of voice biometrics. State-of-the-art solutions in this field are actively addressing the need to develop robust defences against not only traditional security risks but also the voice spoofing and deepfake technologies. As the use of voice biometrics continues to expand in applications like access control, financial transactions, and identity verification, it is essential to address these scientific challenges and opportunities to ensure the trustworthiness and reliability of voice-based authentication methods.

SCIENTIFIC CHALLENGES

- Developing AI models to enhance the security and trustworthiness of voice biometrics is a complicated task. This requires the creation of superior algorithm capable of distinguishing real from synthetic voices. This challenge requires the combination of cutting-edge deep learning techniques with voice recognition to continuously adapt and secure against emerging threats. This involves dealing with various accent and language variations, background noise, and voice quality issues.
- Protecting sensitive voice data is another crucial aspect. This challenge involves developing mechanisms that safeguard stored and transmitted voice samples. It needs a deep understanding of data encryption and secure communication protocols designed to voice biometrics.



- Addressing voice spoofing is important because it directly impacts the reliability and security of voice biometric systems. As voice authentication becomes more common in various sectors, including finance, healthcare, and access control, the threat of voice spoofing presents a significant risk. Developing robust anti-spoofing techniques is necessary to ensure the trustworthiness and integrity of voice-based security measures, maintaining user confidence in the technology and safeguarding sensitive information against fraudulent activities. This challenge needs advanced signal processing, machine learning, and behavioural analysis methods.
- Advancing techniques for the detection of fake voice samples requires exploring speech characteristics and signal analysis. This challenge requires not only identifying synthetic voice attributes but also understanding how these attributes differ from natural human speech. This challenge needs deep learning, feature engineering, and acoustic analysis to design more accurate and reliable fake voice detection methods.

Research activities

- Develop AI models for secure voice biometrics, integrating encryption, privacy-preserving methods, and fake voice detection.
- Investigate methods for detecting and preventing voice spoofing, as well as the generation of fake voice samples.
- Perform accurate testing to ensure the system's trustworthiness, reliability, and fake voice detection capabilities.

Expected results

- At least one scientific publication (conference paper)
- Expected results involve the investigation of novel methods to detect and prevent voice spoofing, ensuring the system's strength against manipulated voice samples. Moreover, accurate testing will be conducted to verify the system's trustworthiness, reliability, and its capabilities in detecting fake voices.

Possible Host Organisations

- BME (Géza Németh)

Vertical Space

VS.1 Synthetic dataset generation of foreign object debris on runways and FATOs

Keywords: Synthetic Dataset; Image Recognition; FOD; Urban Air Mobility.

STATE-OF-THE-ART

Foreign Object Debris (FOD) found in runways and Final Approach and Takeoff Areas (FATOs) pose a major risk to airport, heliport and vertiport operations. To detect and mitigate these hazards, it is essential to periodically screen landing surfaces. Several commercial solutions are available that utilize various combinations of ground-based sensors (e.g., radar) to detect hazardous objects and inform human operators of their locations. In this context, image recognition models based on machine learning techniques have the potential to enhance the accuracy and effectiveness of FOD detection systems, facilitating autonomous operations and onboard FOD detection for unmanned and autonomous aircraft. However, while object recognition is a well-established application of machine learning, specific domain datasets are required to train, test, and evaluate these systems. Traditionally, the collection and curation of high-quality datasets is costly, time-consuming, and labour-intensive.

SCIENTIFIC CHALLENGES

The primary challenge is to build a robust dataset generation tool and pipeline utilizing 3D simulation technologies to produce a comprehensive FOD dataset that must enable the training of image recognition models across a diverse range of weather conditions. The resulting synthetic datasets must be evaluated and compared against real world images, as part of the challenge lies in bridging the gap between synthetic images and real-world images and to evaluate the transferability of image recognition models trained with synthetic datasets.

RESEARCH ACTIVITIES

The selected candidate is expected to:

- Assist with the development of a realistic 3D visual simulation tool based on Unreal Engine.
- With the help of the mentioned tool, generate synthetic datasets of both static and moving FOD on runway surfaces and FATOs under several weather conditions.
- Train and evaluate machine learning image detection models with the generated dataset and assess their effectiveness with real world images.

EXPECTED RESULTS

The main expected result is the generation of a synthetic FOD dataset suitable for training AI image detection and classification models, in such a way the AI models trained with it would be able to bridge the gap in performance between synthetic images and real-world images. The beneficiary is expected to disseminate their results to the ENFIELD consortium on, at least, one of the planned workshops and to produce a technical report describing the methodology or tool being developed during the exchange. The results of this exchange may lead to a per-reviewed publication on a journal or conference.

POSSIBLE HOST ORGANISATIONS

- BAS (Ignacio Vidal and Enrique Casado)

Vertical Space

VS.2 Detection of potential water illegal abstractions using Artificial Intelligence and Earth Observation
Keywords: Agriculture; Water Resources; Artificial Intelligence; Earth Observation; Climate Crisis.
STATE-OF-THE-ART
Remote Sensing and Earth Observation are widely used in irrigation management in order to retrieve useful information about the cultivated crop types and their current crop growth stage. By combining this with evapotranspiration rates, the amount of irrigation needs can be calculated. Furthermore, by monitoring the spatiotemporal patterns of soil moisture using spaceborne sensors the detection of irrigation events can be determined. The employment of Remote Sensing for the detection and monitoring of potential illegal abstractions is still unexplored. Still, a recent research work showed that by monitoring the vegetation health using satellite vegetation indices and by observing unusual patterns of healthy vegetation can be beneficial for the detection of potential illegal water abstractions or illegal irrigation activities.
SCIENTIFIC CHALLENGES
Agriculture industry around the world suffers from Climate Change. Water scarce areas are suffering from irrigation water shortage due to drought events. Moreover, in areas like Cyprus and the rest of Mediterranean basin is noticed that farmers who are cultivating rainfed crops and not only are illegally using more water for irrigation purposes. The investigation of this research challenge lacks solutions.
RESEARCH ACTIVITIES
• Data collection. • Literature review • Development of AI model for detection of illegal water abstractions. • Preparation of scientific manuscript for journal publication.
EXPECTED RESULTS
• 1 journal publication submitted. • 1 model/algorithm/software/framework.
POSSIBLE HOST ORGANISATIONS
• ECoE (Michalis Mavrovouniotis)

Vertical Space

VS.3 Causal Machine Learning model to identify agricultural practices aiding in yield productivity improvement using Earth Observation (EO) data.
Keywords: Machine Learning; Earth Observation; Agriculture; Food Security; Big Data; Artificial Intelligence.
STATE-OF-THE-ART
Industrial applications utilizing EO data for yield prediction and estimation in order to support farmers are arising. Causal Machine Learning (CML) is still undiscovered withing the discipline of Earth Sciences. CML can help through its capabilities to explore a problem further than correlations of data to a problem by detecting the cause and estimating the cause's effect. Causal Inference and Causal Analysis by employing Double Machine Learning (DML) and other more traditional methodologies are widely used in the sectors of economics and business intelligence.
SCIENTIFIC CHALLENGES
Personalized applications dedicated to farmer's practices in combination with EO data are still an undiscovered path. Farmers must be in the centre to help them improve their yield productivity by identifying the cause effect of the different agricultural practices (e.g., irrigation management, fertilizations) to yield productivity. The combination of causal machine learning on EO is still a not quite explored path.
RESEARCH ACTIVITIES
• Data collection. • Literature review • Development of AI model for yield prediction. • Application of Causal Machine Learning to identify the effect of agricultural practices or environmental conditions to yield. • Preparation of scientific manuscript for journal publication.
EXPECTED RESULTS
• 1 journal publication submitted. • 1 model/algorithm/software/framework
POSSIBLE HOST ORGANISATIONS
• ECoE (Michalis Mavrovouniotis)

Vertical Manufacturing

VM.1 Context-agnostic Computer Vision human detection

Machine Learning; Deep Learning; Computer Vision; Domain Shift; Industry 5.0.

STATE-OF-THE-ART

Current industrial trends highlight the need of systems aimed at supporting the manufacturing operators in their daily tasks. These systems, in particular for what concerns the electro-mechanical ones (e.g., collaborative robots) need specific safety measures to be hosted in the same workplaces the operators perform their activities in. Recent advancements in the digitalisation of shopfloor, as well in the Computer Vision (CV) domain are allowing a safer environment where the operators are able to increase their productivity, and to be less impacted by hazardous operations. The most promising implementations in this field have been mainly studied in laboratory alike environments, while the integration into real industrial scenarios needed further adaptation for the CV algorithms, usually based on Machine Learning (ML) models (phaenomenon known as "domain shift problem"). To address this challenge, developing a domain-agnostic generalized visual controller is essential to ensure safety in human-robot collaboration.

SCIENTIFIC CHALLENGES

The hosting institution will make available an industrial setup close to the one depicted in the aforementioned work1, to allow the winner to start from an operational environment embodying the base already available in the literature. Besides of this, the proposal is supposed to address the following objectives:

- Development of a robust framework for human detection and interaction recognition.
- Bridging the domain gap between training and inference time.
- Enhancing safety in human-centric operations in industrial environment.
- Improvement of performances from existing sources in literature.

RESEARCH ACTIVITIES

The candidate is expected to have a good knowledge of Machine Learning and Deep Learning applications in Computer Vision. Moreover, the following steps should be taken:

- Investigating existing approaches and proposing a novel framework for visual human-robot/human-machine interaction (HRI) detection and classification.
- Analysing the proposed framework's robustness against domain shift.
- Design and implement algorithms that improve and reduce the effect of the visual domain gap and create a visual domain agnostic model.

EXPECTED RESULTS

The expected results for this work should enhance human-robot/human-machine interaction safety by achieving visual distribution agnosticism during inference. By doing so, the visually automated HRI models should exhibit minimal performance degradation when deployed in new environments. This means the model can handle the same task even if the objects and robot's visual appearance differ from the training data. The beneficiary is expected to develop a demo for such a task and to produce at least one journal peer-reviewed publication (publication plan is expected and will be evaluated against its ambition and feasibility).

POSSIBLE HOST ORGANISATIONS

- POLIMI

VM.2 Machine Learning-based stress detection for human operators

Keywords: Machine Learning; Industry 5.0; Human-Centric Manufacturing; Signal Processing; Privacy.

STATE-OF-THE-ART

The current phenomenon of ageing of European population is supposed to heavily affect the manufacturing environment, whose operators will (on average) not be employable anymore in repetitive or heavy physical tasks. Current mitigation actions include the usage of empowering technologies (exoskeletons, collaborative robots) which are however increasing the lead time for operations. The need for measures that make these systems enabled only in conditions of physical/mental/emotional stress of the operator is hence supposed to fairly match the wellbeing of the operators and the overall performances of the production system.

SCIENTIFIC CHALLENGES

The hosting institution will make available equipment made by different wearable sensors (EEGs, ECGs, EMGs, eye-trackers, EDA, Skin Temperature (ST), cameras) open to be included or not in the data sources. The winner of the open call is then supposed to create a model able to detect the operators' stress from the aforementioned sensors and systems. Integration support will be provided by the host too. Through these, the candidate is then supposed to develop his/her research and implementation, which will consist in:

- Knowledge extraction about the physical, mental, or emotional stress from the data available
- Addressing the privacy-related issues in framework of the EU regulations.
- Demonstration in laboratory environment with real-like manufacturing tasks (e.g., assembly, kitting, sorting…)

RESEARCH ACTIVITIES

The candidate is expected to have a good knowledge of Data Analysis and Processing Methods. Moreover, the following steps should be taken:

- Investigating existing approaches and proposing a novel framework for physical, mental, or emotional stress detection and classification.
- Design and implement algorithms able to detect physical, mental, or emotional stress in a person involved into manufacturing operations.
- Analysing the proposed framework's robustness against change of the operator (e.g., sex and age).

EXPECTED RESULTS

The expected results for this work should enhance wellbeing of operators in manufacturing environment, as well as the work is supposed to contribute to the applied research activities aimed at extending the European population's employability.
The developed model is supposed to be made public with Open Access rights.
Anonymized dataset is also supposed to be published in the interest of the broader research community as a matter of results' verifications. The beneficiary is expected to develop a demo for such a task and to produce at least one peer-reviewed journal publication (publication plan is expected and will be evaluated against its ambition and feasibility).

POSSIBLE HOST ORGANISATIONS

- POLIMI